Sustainable Energy Technologies and Assessments 84 (2025) 104702

Contents lists available at ScienceDirect

Sustainable Energy Technologies and Assessments

journal homepage: www.elsevier.com/locate/seta

Cyber-attack resilient data-driven approaches for early fault prediction system for wind turbines☆

Animesh Sarkar Tusher [a], Md. Abdur Rahman [b,*], Md. Rashidul Islam [a,c], Md. Arafat Hossain [a], Adnan Anwar [b], M.J. Hossain [c]

[a] Department of Electrical and Electronic Engineering, Rajshahi University of Engineering & Technology, Kazla, Rajshahi 6204, Bangladesh
[b] School of Information Technology, Deakin University, Geelong, VIC 3216, Australia
[c] School of Electrical and Data Engineering, University of Technology Sydney, Ultimo, NSW 2007, Australia

ARTICLE INFO

Dataset link: https://www.kaggle.com/datasets/wasuratme96/iiot-data-of-wind-turbine/

Keywords:
Predictive maintenance (pdM)
Wind turbine (WT)
Early fault prediction
False data injection attacks (FDIAs)
Adversarial training

ABSTRACT

Predictive maintenance (PdM) technologies, facilitated by smart sensors and artificial intelligence, are increasingly adopted in smart grids to ensure reliable operation and prevent financial losses and physical damage. However, these data-driven systems remain susceptible to cyber-attacks. This study examines the resilience of machine learning and deep learning models in predicting wind turbine (WT) faults 30 min in advance under False Data Injection Attacks (FDIAs). Random Forest (RF) and eXtreme Gradient Boosting (XGBoost) outperformed Decision Tree Classifier (DTC) and Gradient Boosting Machine (GBM) as fault predictors under normal conditions. Consequently, a comprehensive evaluation was performed using RF, XGBoost, and Long Short-Term Memory (LSTM) models under cyber-attacks, revealing that such attacks can reduce prediction accuracy by up to 15% and recall by 28%. Both autoencoder-enabled LSTM and adversarial training were implemented as defense mechanisms. While the autoencoder improved stability, adversarial training achieved superior robustness, making XGBoost the most resilient model, maintaining reliable fault prediction under cyber threats with almost no loss. Unlike existing studies focusing solely on fault prediction with clean SCADA data, this work integrates cyber-attack resilience assessment, adversarial defense, and structured data processing into a unified framework, bridging the gap between reliability, security, and data quality for trustworthy WT PdM.

Introduction

Predictive Maintenance (PdM) is one of the key solutions in Industry 4.0 [1], which employs sophisticated machine learning (ML) and Internet of Things (IoT) devices to make certain the faults in components and systems can be predicted. New developments in IoT sensors and ML algorithms have led to a broader adoption of PdM across many industries, especially in wind energy. Furthermore, wind turbine systems can leverage IoT sensors, ML, artificial intelligence, digital twins, and other emerging technologies to enable real-time condition monitoring, early fault prediction, and health management, thereby reducing maintenance costs and enhancing system reliability [2]. With the potential of saving up to 60% of service and maintenance costs [3], the popularity of IoT and ML-enabled PdM is on the rise. However, both IoT devices and ML models are susceptible to cyber-attacks [4], which may seriously threaten the integrity and reliability of PdM systems. Cyber-attacks against PdM systems in wind farms can result in catastrophic outages, degradation of physical assets, and power grid instability. For instance, a denial-of-service attack exploited the vulnerability of the communication network between the control center and wind generation sites in Utah, USA, in March 2019 [5].

As the total installed wind power capacity continues to grow, controlling PdM costs becomes more crucial, since it accounts for around 10%–15% and 20%–25% of onshore and offshore wind farm operational costs, respectively [6]. Simultaneously, implementing effective protective strategies is critical for enhancing Wind Turbine (WT) reliability and minimizing downtime, which ensures optimal efficiency of wind power generation [7]. Consequently, with the rising awareness of such needs, numerous fault prediction models have been introduced for WTs. For instance, [8] proposes a probabilistic WT fault

☆ This work investigates the vulnerability of wind turbine predictive maintenance systems to cyber-attacks and proposes adversarial training to enhance the robustness of machine learning models.

* Corresponding author.
E-mail addresses: eeengg.animesh@gmail.com (A.S. Tusher), md.a.rahman@deakin.edu.au (M.A. Rahman), rashidul@eee.ruet.ac.bd (M.R. Islam), 1601037@student.ruet.ac.bd (M.A. Hossain), adnan.anwar@deakin.edu.au (A. Anwar), jahangir.hossain@uts.edu.au (M.J. Hossain).

https://doi.org/10.1016/j.seta.2025.104702
Received 17 August 2025; Received in revised form 27 October 2025; Accepted 13 November 2025
Available online 14 November 2025

prognosis scheme utilizing a multivariate time series-based mutual information estimator and confidence calibration to enhance fault probability estimation. However, combining multiple computationally intensive components, such as the mutual information estimator, genetic algorithm, ensemble temperature scaling, and Long Short-Term Memory (LSTM) network, makes this approach resource-demanding. Similarly, [9] develops a genetic algorithm-based ensemble model integrating Random Forest (RF), Extra Trees, and XGBoost for fault detection using SCADA data, yet its static threshold-based decision process is prone to false alarms under noisy or manipulated inputs. In another work, [10] introduces an interpretable convolutional temporal–spatial attention network that detects transitions from normal to early fault stages through nonlinear modeling, while [11,12] employ complex hybrid architectures—such as CNN–Vision Transformer and CNN–BiGRU–attention autoencoder—which, despite their high accuracy, significantly increase model complexity and computational burden. Moreover, [13] combines RF with LSTM for short-term fault prediction but suffers from low precision and recall, limiting its practical applicability.

In addition, several other works have focused primarily on WT fault detection and classification rather than early fault prediction or PdM. For example, [14] introduces a deep residual network that integrates convolutional residual blocks with squeeze-and-excitation units and a new activation function (1Dmeta-ACON) to enhance feature extraction. Similarly, [15] utilizes nonstationary vibration signal analysis with conventional ML classifiers for blade fault detection, while [16] combines traditional ML algorithms with neural networks for general WT fault classification. Although these works [14–16] achieve effective fault identification, they primarily emphasize post-fault detection and lack mechanisms for proactive maintenance decision-making.

Furthermore, some studies have concentrated on specific WT components rather than holistic WT-level PdM. For instance, [17] develops a hybrid ResNet50–Xception model for early blade fault detection, [18] designs a transformer-based predictor for IGBT module failures, and [19] proposes an unsupervised dynamic graph representation framework with spatial–temporal self-attention for gearbox and generator faults. Similarly, [20] introduces a transferable incipient fault detection scheme using Average Integrated Power Spectral Density (AIPSD) with One-Class SVM, while [21,22] limit their analyses to gearbox fault prediction. Lastly, [23] focuses solely on the pitch system. Collectively, these studies [17–23] demonstrate promising component-level insights. However, they lack generalizability for early fault detection across the entire turbine system, underscoring the need for more unified, computationally efficient, and noise-resilient PdM frameworks.

PdM appears to offer significant potential benefits in terms of WT's reliability improvements, but its reliance on IoT and Artificial Intelligence (AI) (i.e., ML & DL)-based technologies makes it vulnerable to cyber-attacks. In interconnected systems, attackers can intrude on communication channels and inject falsified sensor data to manipulate operational states without triggering alarms, as evidenced in DC microgrids, where False Data Injection Attacks (FDIAs) have been shown to induce overcurrent or overvoltage conditions covertly [24]. These stealthy attacks are challenging to detect using traditional model-based or AI-based techniques, as both often suffer from model uncertainty, computational complexity, and threshold-design limitations [24]. Such vulnerabilities highlight the risks of cyber–physical compromise driven by financial or disruptive motives, as attackers target system integrity, confidentiality, and authenticity [25]. Historical incidents, such as the Ukraine power grid cyber-attack that left 230,000 people without electricity [26], further underscore the potential consequences. Despite extensive research on early fault prediction for WTs, the impact of FDIA and other cyber-attacks on PdM systems remains largely unexplored to date, which highlights the necessity of this work. Hence, this study aims to address this gap by evaluating the resilience of ML and DL-based WT's early fault prediction systems under cyber-attack conditions, specifically FDIAs.

Another key challenge in the PdM of WTs is identified in [27], which is caused by SCADA data usage. While SCADA data offers a cost-effective solution for condition monitoring, requiring minimal to no additional investment from wind farm operators [6], it presents significant challenges due to its large volume, random variability, and susceptibility to external environmental influences. The data often contains missing or erroneous values, and its temporal dependencies require advanced algorithms for effective processing. Despite extensive use of SCADA data in prior works [8–23], none have provided a systematic or formal approach to preprocess and refine SCADA data specifically for early fault prediction in WTs. Therefore, to the best of authors knowledge, the following key research problems (RPs) remain largely unresolved and demand further investigation.

(RP1) Existing studies have validated their proposed approaches using only clean and reliable SCADA datasets. However, it remains unexplored whether these methods can maintain consistent performance under cyber-attack scenarios or in the presence of maliciously manipulated data.

(RP2) If AI-based diagnostic and prediction models are proven to be vulnerable to cyber-attacks, it is essential to investigate and develop effective defensive mechanisms capable of mitigating the associated security risks and preserving model reliability. However, this particular sector for WTs' PdM applications has remained unexplored, to the best of the authors' knowledge.

(RP3) Although SCADA data serve as the foundation for WT fault diagnosis, the systematic preprocessing of raw SCADA data for early fault detection remains underexplored. Therefore, identifying a robust and formalized data preprocessing strategy is crucial to ensure data quality, feature consistency, and model readiness for WTs' PdM applications.

Hence, to address the above-mentioned RPs, this work investigates the vulnerability of DL and ML-based WT's early fault prediction models to cyber-attacks through a comparative analysis while proposing a countermeasure. This work also demonstrates a comprehensive systematic approach that integrates multiple steps, including data integration, fault status assignment, early fault prediction target preparation, handling class imbalance, and feature selection. It ensures that the SCADA data is preprocessed effectively for robust predictive maintenance and cyber-resilient AI model training. In summary, this paper makes the following contributions.

1. The impact of FDIAs is investigated on PdM performance for WTs using state-of-the-art ML and DL models under both single and mixed-attack scenarios. Findings reveal that scaling attacks are the most severe, leading to significant performance degradation, with DL models being the most vulnerable.
2. This work develops defense mechanisms, including an autoencoder-enabled LSTM and an adversarial training-based approach, to enhance PdM system resilience against FDIAs. Findings show that the adversarially trained (AT) XGBoost model exhibits the highest robustness among all evaluated ML and DL models.
3. A comprehensive approach to processing raw SCADA data is presented, including steps like data integration, fault status labeling, early fault prediction target creation, class imbalance handling, and feature selection. This offers valuable insights for enhancing WT's PdM systems.

The remainder of this paper is organized as follows. Section c presents a detailed overview of cyber-attack models and countermeasure techniques. Section 18 describes data processing, AI-based PdM models, and experimental analysis. Finally, Section 18 concludes the study with key findings and future research directions.

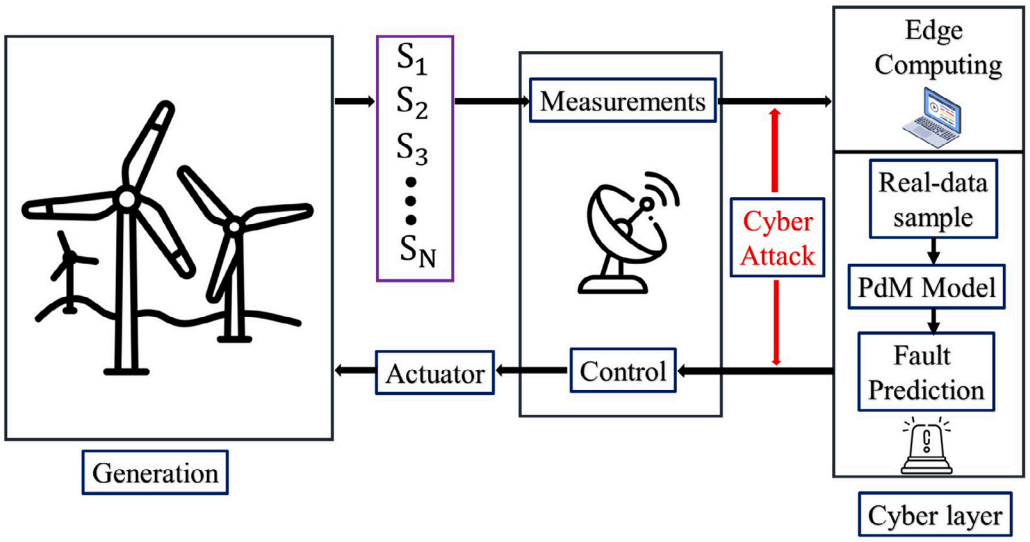

Fig. 1. Data Flow and Cyber-Attack Points in a Wind Turbine PdM System.

Methodologies

A wind farm is essentially a cyber–physical infrastructure that includes interconnected subsystems in both the cyber and physical layers, equipped with real-time data transmission, feedback capabilities, and enhanced accessibility, thereby improving the application of ML methods for real-time PdM of WTs. Modern wind farms increasingly incorporate advanced digital technologies, such as SCADA communication protocols (such as Modbus TCP/IP, DNP3, or IEC 60870-5-104) for data acquisition and transmission in wind farms [9,12,16], which enables remote monitoring of WT's faults, automation, and control of wind energy production and integration into the electric grid. This digitalization expands the cyber layer of wind farms, inherently increasing their cyber vulnerabilities. The seamless integration of physical and cyber layers is designed to enable real-time condition monitoring of WTs, thereby facilitating optimal decision-making for PdM across the system's lifecycle.

Fig. 1 illustrates a framework for understanding cyber-attack vulnerabilities in wind farms, where physical components interact with a cyber layer to enable predictive maintenance. Sensors $(S_1, S_2, S_3, \ldots, S_N)$ on wind turbines collect critical data, including measurements of temperature, vibration, and other operational parameters. This data is then sent wirelessly to local monitoring systems, which in turn regulates turbine operations by issuing control commands via actuators. This data also feeds into an early fault prediction system, which anticipates equipment failures, enhancing maintenance and reducing downtime.

However, this data flow introduces vulnerabilities. As indicated in the cyber-attack pathway in Fig. 1, cyber-attacks can manipulate sensor data or disrupt control commands, potentially causing malfunctions or equipment failure. These attacks undermine the accuracy of the early fault prediction system by either masking real faults or generating false alerts, diminishing the reliability of PdM. The monitoring component, used by operators to oversee turbine status, is also susceptible to unauthorized access, allowing attackers to interfere with alarms and delay responses to actual faults [5]. To counter these threats, it is crucial to implement cybersecurity measures to enable robustness against abnormal patterns in PdM systems.

Attack modeling

The use of data and wireless communication in PdM systems for WTs has considerably increased their vulnerability to potential cyber-attacks. This paper focuses on analyzing the impact of cyber-attacks targeting the wind speed measuring sensor data. Specifically, it evaluates how cyber-attacks affect the system's ability to predict WT faults 30 min in advance, while monitoring is performed at the cyber layer. Among different cyber-attacks to which the smart grid is vulnerable, the study in [28] highlighted FDIAs as the most common attack, while the study in [5] discussed the ability of FDIAs to manipulate sensor data, disrupt control commands, mislead PdM systems, and unauthorized access to maintenance systems in wind firms. Therefore, the objective of this paper is to assess the impact of FDIAs on the WT's early fault prediction systems and propose a corresponding solution.

Let the measurement vector of sensors is, $\boldsymbol{S}(\boldsymbol{x}) = [\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_N] \in \mathbb{R}^{1\times N}$ in digitally enabled wind firms. Then, the information is transmitted by a transmitter via the wireless communication medium, as expressed in Eq. (1).

$$\mathbf{z}(\mathbf{x}) = \boldsymbol{S}(\boldsymbol{x}) + \boldsymbol{e} \tag{1}$$

where **e** is the measurement error vector of WT's sensors or noise added by the transmitter and can be expressed as $\boldsymbol{e} = [e_1, e_2, e_3, \ldots, e_N] \in \mathbb{R}^{1\times N}$.

In FDIAs, an intruder gains simultaneous access to a subset of real-time measurements and manipulates them using an attack vector **a** $\in \mathbb{R}$. As a result of the FDIAs, the system's state estimation becomes compromised, leading to an erroneous state, given by Eq. (2).

$$\hat{\mathbf{z}}(\boldsymbol{x}) = \widehat{\boldsymbol{S}}(\boldsymbol{x}) + \boldsymbol{c} \tag{2}$$

Due to the erroneous state, WT fault prediction can be compromised, which is investigated in this work. This study does not focus on developing new FDIA models or cyber-attack templates. The adversary models employed in this research are inspired by those outlined in Ref. [29,30]. Here, this work specifically focuses on scaling, pulse, and their combination to evaluate the impact of FDIA on models trained with clean data, as well as to train a model using both clean and perturbed data (adversarially trained model). Finally, the AT model is subjected to random attack to assess its robustness against unknown FDIAs.

Scaling attack

Let $D \in \mathbb{R}^{M\times N}$ represent the dataset, where M is the number of time steps and N is the number of features. $D_{attack} \in \mathbb{R}^{M\times N}$ is the dataset after the cyber-attack. Scaling attacks simulate data manipulation by altering variables over a specific period through multiplication with a

scaling parameter denoted as λ_S. The mathematical representation of this attack model is given below in Eq. (3).

$$D_{attack} = (1 + \lambda_S) \times D, \quad \text{for } t_s < t < t_e \tag{3}$$

where t_s and t_e represent the beginning and end of the cyber-attack, respectively. This operation is repeated over 40% of the test data in the case of a single template attack or 20% for a mixed attack.

Pulse attack

Pulse attacks cause data contamination by abruptly increasing or decreasing variable values at specific points within the attack duration, utilizing an attack parameter denoted as λ_p. The mathematical formulation of this attack model is presented below in Eq. (4).

$$D_{attack} = (1 + \lambda_p) \times D, \quad \text{for } t = t_p \tag{4}$$

where t_p, D, and D_{attack} represent the duration of a pulse attack, the unaltered original value, and the value compromised by cyber-attacks, respectively. This is also applied over 40% of the test data for the single template case or 20% for the mixed attack, ensuring that scaling and pulse attacks do not overlap. The systematic execution and evaluation of cyber-attacks are shown in Algorithm 1.

Random attack

A random attack modifies actual measurements by incorporating a randomly generated positive value within a predefined range. The function uses an upper bound (a), and a lower bound (b) to determine the range of possible values. The transformation is applied only at specific time instances ($t \in \tau_a$), whereas, for all other instances, the measurement remains unaltered. A random attack can be mathematically expressed as follows in Eq. (5).

$$D_{attack} = \begin{cases} D, & \text{for } t \notin \tau_a, \\ D + \text{rand}(a, b), & \text{for } t \in \tau_a. \end{cases} \tag{5}$$

where D, and D_{attack} represent the unaltered original value, and the value compromised by cyber-attacks, respectively. This is also applied to 40% of the test data, where a and b are chosen as the maximum and minimum values, respectively, from the historical data of the wind speed measuring sensor data channel.

Algorithm 1: Cyber-Attack Simulation on Test Dataset

Input: $\mathcal{D}_{\text{test}}$: Test data, λ_s, λ_p: Attack parameters, p_{test}: Percentage of test data under attack, step size = 5 for scaling, 1 for pulse.

// $f(\cdot)$ represents each of the different models.

Output: Accuracy and Recall of $f(\cdot)$ under different scenarios.

Preprocess $\mathcal{D}_{\text{test}}$;

// Simulate scaling attack

$\mathcal{D}_{\text{scaled}} \leftarrow \mathcal{D}_{\text{test}}$

for $j = 1$ ***to*** $p_{test} \cdot |\mathcal{D}_{test}|$ **do**

  $start \leftarrow$ Random integer$(1, |\mathcal{D}_{\text{test}}| - \text{step size} + 1)$

  for $i = start$ ***to*** $start + step\ size - 1$ **do**

    $\mathcal{D}_{\text{scaled}}[i] \leftarrow \mathcal{D}_{\text{scaled}}[i] \cdot (1 + \lambda_s)$

  end

end

// Simulate pulse attack

$\mathcal{D}_{\text{pulsed}} \leftarrow \mathcal{D}_{\text{test}}$

for $j = 1$ ***to*** $p_{test} \cdot |\mathcal{D}_{test}|$ **do**

  $start \leftarrow$ Random integer$(1, |\mathcal{D}_{\text{test}}| - \text{step size} + 1)$

  while $start \in S_{test}$ **do**

    $start \leftarrow$ Random integer$(1, |\mathcal{D}_{\text{test}}| - \text{step size} + 1)$

  end

  for $i = start$ ***to*** $start + step\ size - 1$ **do**

    $\mathcal{D}_{\text{pulsed}}[i] \leftarrow \mathcal{D}_{\text{pulsed}}[i] \cdot (1 + \lambda_p)$

  end

end

// Combine scaling and pulse attacks

$\mathcal{D}_{\text{mixed}} \leftarrow \mathcal{D}_{\text{scaled}} \cup \mathcal{D}_{\text{pulsed}}$

// Evaluate Models

Accuracy & Recall $\leftarrow f(\cdot) \leftarrow \mathcal{D}_{\text{test}}, \mathcal{D}_{\text{scaled}}, \mathcal{D}_{\text{pulsed}}, \mathcal{D}_{\text{mixed}}$

Countermeasure technique

This subsection details two adopted defenses against FDIAs in PdM of WTs. Firstly, an Autoencoder (AE)-assisted LSTM that mitigates attacks by reconstructing inputs onto the manifold of normal operation before prediction, and secondly, Adversarial Training (AT) that repeatedly exposes the model to adversarial examples and minimizes task loss on these inputs. The decision boundary is reshaped to be locally stable, thereby improving robustness against attacks at test time.

LSTM with autoencoder

This work employs an AE-LSTM pipeline to mitigate FDIAs in time-series sensing. The AE, trained exclusively on clean data, learns a compact manifold of normal operation and reconstructs inputs by projecting them onto this manifold, thereby suppressing out-of-distribution perturbations introduced by FDIAs [28]. The LSTM then consumes the reconstructed (denoised) sequence for prediction or classification, reducing the propagation of corrupted features into downstream decisions. Hence, this work adapts the LSTM architecture from [31] and trains it jointly with an AE using a composite objective that couples sequence reconstruction (normal-manifold fidelity) with the supervised task loss. This end-to-end scheme yields features that preserve nominal dynamics while remaining robust to injected perturbations. The incorporated AE is with a latent dimensionality of 1024, selected via validation to provide sufficient representational capacity for accurate reconstruction while maintaining robustness to distributional perturbations. This latent size offers a practical trade-off between fidelity and regularization, and is further stabilized with early stopping and dropout during training. These design choices aim to mitigate the impact of FDIAs. Overall, the AE functions as a learned pre-filter while the LSTM enforces temporal coherence, yielding a unified detection-and-mitigation mechanism for CPS/SCADA signals.

Adversarial training

To enhance the robustness of ML and DL models for WT's early fault prediction against FDIAs, this work also employs AT. This technique, where a model is trained on both clean and adversarial examples, improves generalization and reduces sensitivity to attacks. During adversarial training, the model learns from both original inputs D and adversarial examples D_{adv} generated with minimal perturbation to affect model performance. This process can be represented by the following Eqs. (6) and (7).

$$f_{adv}^{aux} = Train(D_{adv} + D) \tag{6}$$

$$f = \mathbf{x} \mapsto f_{adv}^{aux}(D) \tag{7}$$

This dual training approach strengthens the model's resilience, although there is a potential risk of reduced accuracy under normal conditions. To make sure that the model works in a wide range of attack situations, a mixed attack template is used that trains on 30% of the whole dataset using both scaling and pulse attacks. This strategy effectively mitigates the impact of various attacks while maintaining acceptable performance with clean data. The entire procedure is demonstrated in the Algorithm 2. Finally, the performance of each model on the test data was assessed using accuracy and recall, calculated based on the confusion matrix [32] as follows in Eq. (8) and in

Eq. (9).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{9}$$

where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.

Algorithm 2: Adversarial Training

Input: $\mathcal{D}_{\text{train}}$: Training data, $\lambda_s = \lambda_p = 0.6$: Attack parameters, p_s: Scaling attack percentage, p_p: Pulse attack percentage, step size = 5 for scaling, 1 for pulse.

// $f^{\text{adv}}(\cdot)$ represents each of the different adversarially trained models.

Output: $f^{\text{adv}}(\cdot)$

Preprocess $\mathcal{D}_{\text{train}}$

// Generate scaling attack samples.

$\mathcal{D}_{\text{adv}} \leftarrow \mathcal{D}_{\text{train}}$

for $j = 1$ ***to*** $p_s \cdot |D_{adv}|$ **do**

  $start \leftarrow$ Random integer$(1, |\mathcal{D}_{\text{adv}}| - \text{step size} + 1)$

  for $i = start$ ***to*** $start + step\ size - 1$ **do**

    $\mathcal{D}_{\text{adv}}[i] \leftarrow \mathcal{D}_{\text{adv}}[i] \cdot (1 + \lambda_s)$

  end

end

// Generate pulse attack samples.

for $j = 1$ ***to*** $p_p \cdot |D_{adv}|$ **do**

  $start \leftarrow$ Random integer$(1, |\mathcal{D}_{\text{adv}}| - \text{step size} + 1)$

  while $start \in S_s$ **do**

    $start \leftarrow$ Random integer$(1, |\mathcal{D}_{\text{adv}}| - \text{step size} + 1)$

  end

  for $i = start$ ***to*** $start + step\ size - 1$ **do**

    $\mathcal{D}_{\text{adv}}[i] \leftarrow \mathcal{D}_{\text{adv}}[i] \cdot (1 + \lambda_p)$

  end

end

// Train adversarially trained models.

$f^{\text{adv}}(\cdot) \leftarrow \text{Train}(D_{\text{adv}})$

Experimental setup and analysis

The following subsections outline the data collection and preprocessing. It details the model training process, including cross-validation and hyperparameter tuning, to ensure optimal predictive performance. This section also explores the effects of FDIAs on the PdM of WTs. Additionally, it analyzes varying attack intensities and assesses the resilience of the proposed AT model, demonstrating its effectiveness against such threats.

Dataset preparation

As identified in [27], how to use SCADA data to accurately predict overall WT faults remains a great challenge. Hence, this work comes to the rescue. A publicly available dataset [33] is used for this work. This dataset originates from a 3 MW direct-drive wind turbine located at a large offshore wind farm in southern Ireland [12]. It comprises 10-minute interval operational records and real-time alarm data collected through the turbine's SCADA system. The dataset spans an 11-month period, from May 2014 to April 2015. Additional information and detailed descriptions of this dataset are available in [9,12], where the same data source has been employed for related research. The SCADA data (S) contains 65 parameters providing information about WT's conditions, and fault data (F) provides information on fault status in 10-minute intervals. Combining them, preparing the target class, and preparing the dataset of 30 min of early fault prediction is expressed by Eq. (10) through Eq. (13).

Table 1
Parameters and sub-parameters for the PdM task of WT.

SCADA parameters	Sub-parameters
Wind speed	x
Temperature	Inverter temperature Rotor temperature Stator temperature Rectifier temperature Fan inverter temperature Transformer temperature
Generator rotational speed	Maximum rotation Minimum rotation
Active power	Maximum power Average available power from wind
Reactive power	Maximum reactive power Average reactive power
Blade angle	x
x - no sub-parameter	

$$\mathcal{D} = S \cup_T F \tag{10}$$

$$\mathcal{D}(t, \text{Fault}) = \begin{cases} \mathcal{D}(t, \text{Fault}) & \text{if } \mathcal{D}(t, \text{Fault}) \neq \text{NaN} \\ NF & \text{if } \mathcal{D}(t, \text{Fault}) = \text{NaN} \end{cases} \tag{11}$$

$$\mathcal{D}_{\text{binary}}(t, \text{Fault}) = \begin{cases} 0 & \text{if } \mathcal{D}(t, \text{Fault}) = NF \\ 1 & \text{if } \mathcal{D}(t, \text{Fault}) \neq NF \end{cases} \tag{12}$$

$$\mathcal{D}^{\text{pdm}}(t, \text{Fault}) = \mathcal{D}_{\text{binary}}(t + 3, \text{Fault}) \tag{13}$$

From Eq. (14), the fault probability is calculated as $p_{\text{Fault}} = 1.1\%$, indicating that $\mathcal{D}_{\text{pdm}}(t, \text{Fault})$ is imbalanced. To address this, this work applies the balancing technique outlined in Eq. (15)–(17). This process yields a balanced dataset like Eq. (16), and the final dataset is denoted as $\mathcal{D}$ in Eq. (17).

$$p_{\text{Fault}} = \frac{\sum_t \mathcal{D}^{\text{pdm}}(t, \text{Fault})}{|T|} \tag{14}$$

$$\mathcal{D}_{\text{new}} = \mathcal{D}_i^{\text{pdm}} + \lambda \cdot (\mathcal{D}_{\text{knn}}^{\text{pdm}} - \mathcal{D}_i^{\text{pdm}}) \tag{15}$$

$$|\mathcal{D}_{\text{Fault}}^{\text{pdm}}| = |\mathcal{D}_{\text{NF}}^{\text{pdm}}| \tag{16}$$

$$\mathcal{D} = \mathcal{D}_{\text{Fault}}^{\text{pdm}} \cup \mathcal{D}_{\text{NF}}^{\text{pdm}} \tag{17}$$

where $\mathcal{D}_i^{\text{pdm}} \in \mathcal{D}_{minority}^{\text{pdm}}$(a fault condition in this work), $\mathcal{D}_{\text{knn}}^{\text{pdm}}$ is one of the k-nearest neighbors of $\mathcal{D}_i^{\text{pdm}}$, and $\lambda \in [0, 1]$ is a random scalar.

As not all features in SCADA data are relevant to generator faults, it is essential to first identify a subset of SCADA-collected features that accurately represent generator operating conditions. Table 1 lists the selected features used specifically for generator fault prediction.

The roles of these parameters mentioned in Table 1 can be expressed as follows:

1. Power curve deviation, indicating potential fault, as expressed in Eq. (18).

$$P_{\text{turbine}} \neq f(v_{\text{wind}}) \tag{18}$$

2. Mismatch in power generation and rotational speed, suggesting generator malfunction, as expressed in Eq. (19).

$$P_{\text{gen}} \neq g(\omega_{\text{gen}}) \tag{19}$$

3. Reactive power anomaly linked to RPM, indicating electrical irregularities, as expressed in Eq. (20).

$$Q_{\text{reactive}} \neq h(\omega_{\text{gen}}) \tag{20}$$

4. Abnormal generator temperature fluctuation, signaling heat dissipation issues, as expressed in Eq. (21).

$$\Delta T_{\text{temp}} \gg \text{Normal fluctuation} \tag{21}$$

5. Shaft torque discrepancy with load, suggesting generator fault, as expressed in Eq. (22).

$$T_{\text{shaft}} \neq i(P_{\text{gen}}) \tag{22}$$

6. Blade angle anomaly, potentially indicating gearbox issues, as expressed in Eq. (23).

$$\theta_{\text{blade}} \neq j(\text{Gearbox status}) \tag{23}$$

In these equations, P_{turbine}, P_{gen}, Q_{reactive}, ΔT_{temp}, T_{shaft}, and θ_{blade} represent observed values, while functions $f(v_{\text{wind}})$, $g(\omega_{\text{gen}})$, $h(\omega_{\text{gen}})$, and $j(\text{Gearbox status})$ define expected behavior under normal operation.

The parameters mentioned in Table 1 are also validated by Eq. (24) through Eq. (27).

Let $D_1, D_2, \ldots, D_n$ represent the individual parameters in the dataset D, and the entire dataset can be represented by Eq. (24).

$$D = \{D_1, D_2, \ldots, D_n\} \tag{24}$$

Let Y represent the binary fault label (0 for no fault, 1 for fault). Now, the correlation between each pair of features D_i and D_j, and between each feature D_i and the fault label Y, is computed using the Pearson correlation coefficient. Hence, the Pearson correlation coefficient between two features D_i and D_j is expressed by Eq. (25).

$$\rho_{D_i,D_j} = \frac{\text{Cov}(D_i, D_j)}{\sigma_{D_i}\sigma_{D_j}} \tag{25}$$

where ρ_{D_i,D_j} is the correlation coefficient between parameters D_i and D_j, $\text{Cov}(D_i, D_j)$ is the covariance between D_i and D_j, and $\sigma_{D_i}, \sigma_{D_j}$ are the standard deviations of D_i and D_j, respectively. Then, for each parameter D_i, the correlation with the fault label Y is calculated by Eq. (26).

$$\rho_{D_i,Y} = \frac{\text{Cov}(D_i, Y)}{\sigma_{D_i}\sigma_Y} \tag{26}$$

where $\rho_{D_i,Y}$ is the correlation coefficient between feature D_i and the fault condition Y, and σ_Y is the standard deviation of the binary fault label. Then, the overall correlation matrix $\mathbf{R}$ for all selected features D_i and the fault label Y is expressed by Eq. (27).

$$\mathbf{R} = \begin{pmatrix} \rho_{D_1,D_1} & \rho_{D_1,D_2} & \cdots & \rho_{D_1,D_n} & \rho_{D_1,Y} \\ \rho_{D_2,D_1} & \rho_{D_2,D_2} & \cdots & \rho_{D_2,D_n} & \rho_{D_2,Y} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho_{D_n,D_1} & \rho_{D_n,D_2} & \cdots & \rho_{D_n,D_n} & \rho_{D_n,Y} \\ \rho_{Y,D_1} & \rho_{Y,D_2} & \cdots & \rho_{Y,D_n} & \rho_{Y,Y} \end{pmatrix} \tag{27}$$

where each element ρ_{D_i,D_j} represents the correlation between features D_i and D_j, and $\rho_{D_i,Y}$ represents the correlation between feature D_i and the fault label Y. The correlation matrix $\mathbf{R}$ is then visualized using a heatmap, and the parameters listed in Table 1 are identified as having the highest correlation with the fault label. These parameters are thus selected as features for WT's early fault prediction, while other parameters in the dataset are excluded.

Model training, cross-validation, and hyperparameter tuning

This study evaluates the performance of multiple ML models, including Decision Tree Classifier (DTC) [34], Random Forest (RF) [29], eXtreme Gradient Boosting (XGBoost) [29], and Gradient Boosting Machine (GBM) [35], along with a DL model, LSTM [13]. The ML models were specifically trained and optimized for this study, while the LSTM model was reproduced from [13] for comparison.

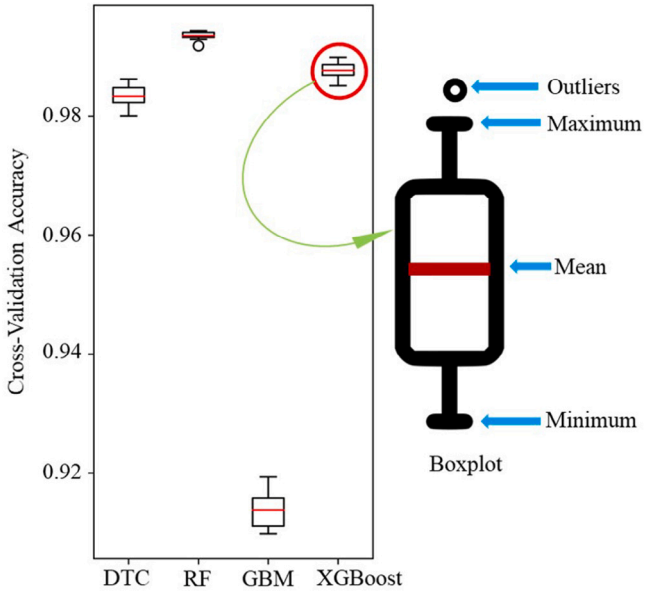

Fig. 2. CVA of ML models along with zoomed-in view of a box plot.

To ensure robust model evaluation, stratified k-fold cross-validation was employed. This technique maintains consistent class distributions across folds, preventing bias due to data imbalance. Given a dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$, where x_i represents the feature vector and y_i denotes the target label, the dataset is partitioned into k folds $D_1, D_2, \ldots, D_k$.

For each fold $t \in \{1, 2, \ldots, k\}$, the training set and validation set are defined as, in Eq. (28).

$$D_{\text{train}}^{(t)} = \bigcup_{\substack{j=1 \\ j \neq t}}^{k} D_j, \quad D_{\text{val}}^{(t)} = D_t. \tag{28}$$

Each model $f_j(\cdot)$ is trained on $D_{\text{train}}^{(t)}$ and evaluated on $D_{\text{val}}^{(t)}$ to compute the accuracy ($\text{Acc}^{(t)}$) for fold t, as expressed in Eq. (29).

$$\text{Acc}^{(t)} = \frac{1}{|D_{\text{val}}^{(t)}|} \sum_{(x_i,y_i) \in D_{\text{val}}^{(t)}} \mathbb{1}\{f_j(x_i) = y_i\}, \tag{29}$$

where $\mathbb{1}\{\cdot\}$ is the indicator function, yielding 1 if $f_j(x_i) = y_i$ and 0 otherwise.

The overall Cross-Validated Accuracy (CVA) for f_j is computed by Eq. (30).

$$\text{CVA}(f_j) = \frac{1}{k} \sum_{t=1}^{k} \text{Acc}^{(t)} \tag{30}$$

This process was carried out for all models, including DTC, RF, XGBoost, and GBM. The CVA from each fold, visualized in the box plot in Fig. 2, demonstrates that RF and XGBoost consistently outperform DTC and GBM in terms of mean accuracy across all folds. Both RF and XGBoost also show narrower interquartile ranges, indicating more stable performance with fewer variations between cross-validation runs. In contrast, GBM exhibits the lowest mean accuracy with a wider spread, suggesting inconsistency and sensitivity to data partitioning. Additionally, a few minor outliers are observed in DTC and GBM, further supporting their lower generalization capability compared to ensemble-based models. Consequently, DTC and GBM were excluded from further steps, such as hyperparameter tuning, cyber-attack impact analysis, and countermeasure application.

To optimize model performance, this study employed randomized search cross-validation (`RandomizedSearchCV`) for XGBoost and RF classifiers, aiming to maximize CVA. The XGBoost model, denoted as $\text{XGB}(\theta)$, where θ represents the hyperparameters, was optimized over a

predefined parameter grid $\mathcal{P}$, as shown in Eq. (31). The values in **bold** represent the selected optimal parameters.

$$\mathcal{P} = \left\{ \begin{array}{ll} n_{\text{estimators}} & \in \{150, 200, \mathbf{250}, 300\}, \\ \text{scale\_pos\_weight} & \in \{5, \mathbf{10}, 15\}, \\ \text{learning\_rate} & \in \{0.1, 0.2, \mathbf{0.3}\}, \\ \gamma & \in \{\mathbf{0}, 3, 5, 7\}, \\ \text{subsample} & \in \{0.8, \mathbf{0.9}, 1.0\} \end{array} \right\} \tag{31}$$

A randomized selection of hyperparameter combinations was evaluated over 10 cross-validation folds to determine the optimal set θ^* that maximized CVA(XGB(θ)), formulated by Eq. (32).

$$\theta^* = \arg\max_{\theta \in \mathcal{P}} \text{CVA}(\text{XGB}(\theta)). \tag{32}$$

Since RF achieved its best performance using default settings, no additional tuning was required for RF.

Experimental analysis

This work evaluates the performance of predictive maintenance models trained on clean data when exposed to different types of cyber-attacks, including scaling, pulse, and mixed attacks (a combination of scaling and pulse perturbations). Each attack type was simulated at three intensity levels, 0.4, 0.6, and 0.8, to analyze the models' resilience under varying degrees of adversarial perturbation. The performance was assessed using accuracy, recall, and the corresponding standard deviations to quantify both predictive capability and stability. To ensure statistical reliability, each experiment was executed ten times under every adversarial condition. Since cyber-attacks can manipulate different data instances across runs, averaging the results and calculating standard deviations provide a more comprehensive and consistent measure of model robustness under diverse operational scenarios.

Performance analysis of normally trained models under different scenarios

Table 2 presents a comprehensive comparison of the performance of two ML models, XGBoost and RF, and one DL model, LSTM, under both normal and cyber-attack conditions. Under normal conditions, all models demonstrated strong predictive capabilities. XGBoost achieved the highest performance with an average accuracy of 0.9896 and recall of 0.9996, closely followed by RF (0.9928 accuracy, 0.9975 recall). The LSTM model performed slightly lower (0.9431 accuracy, 0.9379 recall), indicating that traditional ML models, particularly ensemble-based methods, are better suited for static SCADA feature learning than sequence-based architectures when no adversarial interference is present.

When subjected to scaling attacks, all models experienced a gradual decline in accuracy and recall as attack intensity increased. At 0.4 intensity, XGBoost remained highly resilient (0.9768 accuracy, 0.9769 recall) with the lowest standard deviation (0.000577), demonstrating stable performance across runs. In contrast, LSTM showed a sharp reduction (0.8582 accuracy, 0.8582 recall), indicating higher sensitivity to data distortion. As the attack intensity increased to 0.8, XGBoost maintained strong performance (0.9627 accuracy, 0.9624 recall), while LSTM dropped drastically to 0.8029, confirming its vulnerability to cumulative perturbations.

In the case of pulse attacks, which simulate abrupt and short-duration data disturbances, the performance degradation was more pronounced compared to scaling attacks. At 0.4 intensity, XGBoost and RF both maintained stable performance with accuracy and recall above 0.95, while LSTM declined to 0.8594, showing its limited ability to handle transient anomalies. As the attack intensity increased to 0.8, LSTM's recall fell to 0.8024, whereas XGBoost and RF sustained recall levels of 0.9615 and 0.9255, respectively, indicating their superior resistance to such impulsive data manipulations.

Under mixed attack scenarios (a combination of scaling and pulse perturbations), all models exhibited their lowest overall performance, as this condition introduces both gradual and abrupt data corruptions. XGBoost consistently outperformed others, achieving 0.9613 accuracy and 0.9542 recall at 0.8 intensity, with low standard deviations of 0.0009 and 0.0015, respectively. RF showed slightly higher variation, while LSTM suffered the largest accuracy drop (0.8037) and recall drop (0.8043), reaffirming its limited robustness in hybrid attack environments.

The standard deviation values for both accuracy and recall provide insights into each model's stability and consistency across repeated experiments. A lower standard deviation indicates that the model's performance remains consistent under multiple runs, reflecting robustness to stochastic variations and attack randomness. Conversely, higher deviations, as observed in LSTM, particularly under high-intensity and mixed attack scenarios, imply greater sensitivity to attack-induced noise and unstable learning behavior.

Overall, the analysis confirms that XGBoost demonstrates the highest resilience and consistency across all scenarios, RF performs reliably with slight variability, and LSTM, despite strong performance under normal conditions, remains the most vulnerable to cyber-attacks. The statistical findings emphasize that ensemble-based ML models not only achieve superior predictive accuracy but also maintain greater stability under repeated adversarial conditions.

Performance analysis of autoencoder enabled LSTM under different scenarios

To enhance the resilience of predictive maintenance models against cyber-attacks, this study incorporates an autoencoder-enabled AE-LSTM as a defense mechanism. The autoencoder structure is employed to reconstruct and denoise the input features before passing them to the LSTM network, thereby mitigating the impact of FDIAs and restoring signal integrity. This hybrid approach aims to preserve the temporal learning capability of LSTM while improving robustness to adversarial perturbations.

Table 3 summarizes the performance of the AE-LSTM model under normal and various cyber-attack conditions, including scaling, pulse, and mixed attacks, each tested at three intensity levels (0.4, 0.6, and 0.8). Compared to the normally trained (NT) LSTM, the AE-LSTM demonstrates clear performance improvements across all scenarios. Under normal conditions, the model achieved an average accuracy of 0.9007 and a recall of 0.9008, reflecting stable baseline learning, however, lower than the NT LSTM model. When subjected to attacks, the AE-LSTM maintained relatively consistent performance, with only minor declines in accuracy and recall. For instance, under a 0.8-intensity mixed attack, the most severe condition, the accuracy and recall remained at 0.8885 and 0.8887, respectively, representing a marginal reduction of less than 2% from the normal case.

The standard deviation values for both accuracy and recall remained very low across all tests (ranging between 0.00027 and 0.00085), indicating high model stability and repeatability across ten independent runs. This consistency demonstrates that the autoencoder effectively reduces random variations caused by cyber-attack perturbations. Moreover, the AE-LSTM showed minimal sensitivity to increasing attack intensity, confirming its ability to maintain reliable fault prediction performance even under adversarial conditions.

Although the AE-LSTM demonstrates improved stability and moderate resilience against cyber-attacks, its performance still degrades under high-intensity and mixed attack scenarios. While the reconstruction mechanism helps in partially filtering out injected noise, it cannot fully restore the integrity of corrupted input data. This indicates that the AE-LSTM, despite offering consistent predictions, is not completely effective in mitigating the impact of sophisticated FDIAs. Therefore, to strengthen the system's robustness and enhance resistance to adversarial perturbations, this study introduces adversarial training as the proposed defense mechanism, which is discussed in the next subsection.

Table 2
Performance comparison of normally trained models.

Type	Intensity	Model	Avg. Accuracy	Avg. Recall	Std. of Accuracy	Std. of Recall
Normal	x	LSTM	0.9431	0.9379	–	–
		XGBoost	0.9896	0.9996	–	–
		RF	0.9928	0.9975	–	–
Scaling	0.4	LSTM	0.8582	0.8582	0.001332	0.001339
		XGBoost	0.9768	0.9769	0.000577	0.000574
		RF	0.9611	0.9605	0.004025	0.001504
Pulse		LSTM	0.8594	0.8594	0.000867	0.000859
		XGBoost	0.9761	0.9762	0.000768	0.000766
		RF	0.9571	0.9568	0.001517	0.001543
Mixed		LSTM	0.8585	0.8586	0.002039	0.002017
		XGBoost	0.9766	0.9766	0.000228	0.00027
		RF	0.9474	0.9471	0.018599	0.018789
Scaling	0.6	LSTM	0.8235	0.8236	0.00129	0.001254
		XGBoost	0.9676	0.9653	0.000847	0.026888
		RF	0.9407	0.9457	0.001155	0.000977
Pulse		LSTM	0.8252	0.8254	0.002721	0.002721
		XGBoost	0.9677	0.9676	0.000757	0.000757
		RF	0.9351	0.9516	0.001616	0.010882
Mixed		LSTM	0.8245	0.8248	0.001253	0.001519
		XGBoost	0.9671	0.9669	0.000936	0.000936
		RF	0.9361	0.9355	0.001572	0.001572
Scaling	0.8	LSTM	0.8029	0.8031	0.002062	0.002062
		XGBoost	0.9627	0.9624	0.002982	0.00065
		RF	0.9306	0.991	0.000934	0.00429
Pulse		LSTM	0.8024	0.8023	0.004669	0.004656
		XGBoost	0.9616	0.9615	0.000409	0.000409
		RF	0.926	0.9255	0.001134	0.00115
Mixed		LSTM	0.8037	0.8043	0.001566	0.001997
		XGBoost	0.9613	0.9542	0.000909	0.014717
		RF	0.9268	0.9263	0.001878	0.001846

Table 3
Performance of the Autoencoder-enabled LSTM model.

Type	Intensity	Avg. Accuracy	Avg. Recall	Std. of Accuracy	Std. of Recall
Normal	x	0.9007	0.9008	–	–
Scaling	0.4	0.8953	0.8954	0.000396	0.000396
Pulse		0.8949	0.8949	0.00027	0.00027
Mixed		0.8944	0.8954	0.000447	0.000404
Scaling	0.6	0.8912	0.8913	0.000844	0.000844
Pulse		0.8907	0.8908	0.00085	0.00085
Mixed		0.8910	0.8911	0.000713	0.000713
Scaling	0.8	0.8881	0.8882	0.000844	0.000844
Pulse		0.8875	0.8876	0.000321	0.000321
Mixed		0.8885	0.8887	0.00023	0.000274

Performance analysis of adversarially trained models under different scenarios

The vulnerability of ML models to cyber-attacks, such as data perturbations due to scaling and pulse attacks, emphasizes the need to strengthen model robustness to ensure reliable performance in real-world applications. WTs and other critical infrastructure that rely on AI-based predictive maintenance systems can be severely compromised by even minor manipulations in sensor data, leading to misclassifications, delays in maintenance, or even system failures. The impact on LSTM highlights the susceptibility of DL models that depend heavily on sequential data patterns, while the impact on RF and XGBoost demonstrates that even ensemble methods are not immune to such threats. Thus, mitigating the influence of cyber-attacks on these systems is crucial to ensuring safe and reliable operations, particularly in industries where decision-making is time-sensitive and critical to maintaining operational stability. To address these vulnerabilities, adversarial training provides a promising defense mechanism by incorporating adversarial examples into the training process.

Table 4 presents the comparative performance of AT models, LSTM, XGBoost, and RF, under both normal and attack conditions. Overall, AT models demonstrated significant robustness across all attack scenarios and intensities. Under normal conditions, all models maintained performance levels comparable to their non-adversarial counterparts, confirming that the inclusion of adversarial training does not compromise baseline predictive accuracy. RF achieved the highest accuracy (0.9928) and recall (0.9976), followed closely by XGBoost (0.9897 accuracy, 0.9996 recall) and LSTM (0.9432 accuracy, 0.9380 recall).

When subjected to scaling attacks, all AT models sustained strong predictive performance with minimal degradation. At 0.8 intensity, the AT-XGBoost model achieved 0.9901 accuracy and 0.9902 recall, maintaining nearly identical results to the clean data condition. LSTM and RF also showed improved resistance compared to their NT versions. LSTM achieved 0.9368 accuracy and 0.9369 recall, reflecting a significant improvement over the AE-LSTM model, which experienced noticeable accuracy drops under similar conditions. The low standard deviations observed across all models (less than 0.001 in most cases) indicate high stability and consistent results across ten repeated experiments, reinforcing the effectiveness of adversarial training in producing resilient model performance.

Under pulse and mixed attack scenarios, AT models continued to exhibit superior resilience. AT-XGBoost maintained recall values above 0.96 even under the most intense mixed attacks (0.8), while AT-RF and AT-LSTM remained nearly 0.93, outperforming both their non-adversarial and autoencoder-enhanced variant for LSTM. The robustness of XGBoost is attributed to its strong gradient-boosting framework, which effectively captures non-linear data interactions while adapting to perturbation patterns learned during adversarial training.

Compared to the AE-LSTM model, the AT models—particularly AT-XGBoost—exhibited more consistent robustness across all attack intensities. While the AE-LSTM reduced variability and enhanced stability, its ability to fully counteract high-intensity FDIAs was limited. In contrast, AT models not only maintained high predictive accuracy but also demonstrated improved generalization against unseen perturbations, suggesting superior adaptation to attack conditions.

Table 4
Performance comparison of adversarially trained models.

Type	Intensity	Model	Avg. Accuracy	Avg. Recall	Std. of Accuracy	Std. of Recall
Normal	x	LSTM	0.9432	0.9380	–	–
		XGBoost	0.9897	0.9996	–	–
		RF	0.9928	0.9976	–	–
Scaling	0.4	LSTM	0.9395	0.9396	0.0001924	0.000228
		XGBoost	0.9901	0.9901	0.000239	0.000239
		RF	0.992	0.9921	0.00007	0.00007
Pulse		LSTM	0.9398	0.9398	0.000477	1.000477
		XGBoost	0.9772	0.9771	0.000664	0.000634
		RF	0.9567	0.9565	0.001824	0.001903
Mixed		LSTM	0.9395	0.9395	0.000605	0.000626
		XGBoost	0.9771	0.9771	0.000241	0.000223
		RF	0.9597	0.9594	0.000751	0.000757
Scaling	0.6	LSTM	0.9387	0.9387	0.0003899	0.0003535
		XGBoost	0.99	0.9901	0.000114	0.000158
		RF	0.9922	0.9922	0.00007	0.00005
Pulse		LSTM	0.9387	0.9387	0.000179	0.000179
		XGBoost	0.9675	0.9674	0.000356	0.000358
		RF	0.9423	0.9419	0.014186	0.014364
Mixed		LSTM	0.9383	0.9384	0.000297	0.000288
		XGBoost	0.9681	0.9679	0.000646	0.000594
		RF	0.9416	0.9411	0.001616	0.001602
Scaling	0.8	LSTM	0.9368	0.9369	0.0007463	0.000737
		XGBoost	0.9901	0.9902	0.00023	0.000228
		RF	0.9923	0.9923	0.000158	0.000167
Pulse		LSTM	0.9373	0.9373	0.000356	0.000356
		XGBoost	0.9612	0.9611	0.001088	0.0010876
		RF	0.9286	0.9281	0.000694	0.0007325
Mixed		LSTM	0.9367	0.9368	0.000472	0.000492
		XGBoost	0.9609	0.9608	0.001031	0.001031
		RF	0.9316	0.9311	0.001111	0.001112

Hence, adversarial training proves to be a highly effective defense mechanism, significantly enhancing model robustness against cyber threats. Among the three models, XGBoost emerges as the best-performing model, consistently achieving the highest accuracy and recall across all attack scenarios. RF also demonstrates strong resilience, with only minor fluctuations in recall under mixed attack conditions. While LSTM shows notable improvements with adversarial training, it still lags slightly behind XGBoost and RF. Overall, XGBoost is the most robust and reliable model for handling FDIAs, making it the preferred choice for deployment in security-critical environments. XGBoost also demonstrates its superiority in previous studies, such as [36], where it proves to be more robust to noise and variations in SCADA data compared to LSTM. In scenarios involving higher temperature variations, LSTM's predictions show larger errors, highlighting its limitations. The underlying advantage of XGBoost lies in its use of recursive partitioning within an ensemble framework, which simplifies the training process and minimizes optimization complexity, especially when compared to the multi-layer architecture and extensive parameter space of LSTM.

To further evaluate the generalization ability of AT models, they are exposed to random attack, a perturbation type unseen during training. Despite this challenge, the models demonstrate remarkable robustness, reinforcing the effectiveness of adversarial training in enhancing resilience against unforeseen threats. As illustrated in Fig. 3, XGBoost and RF maintain exceptionally high accuracy and recall, both exceeding 0.98, indicating strong adaptability to new attack patterns. LSTM, although performing slightly lower, still retains accuracy above 0.90, showing substantial improvement over its normally trained version. The minimal performance gap between accuracy and recall for all models further confirms balanced classification behavior and stable learning under adversarial influence. This confirms that adversarial training not only strengthens models against known attacks but also equips them with the capability to withstand novel adversarial conditions, making XGBoost and RF the most reliable choices for real-world cybersecurity applications.

Conclusion

The early fault prediction is crucial for the safe and efficient operation of WT, as it can assist in avoiding significant economic losses and even accidents. However, the emerging threat of cyber-attacks can significantly enhance the failure risks of ML- and DL-based approaches dedicated to PdM. Therefore, this study addresses this issue and shows that ML models, especially RF and XGBoost, are better at accuracy and durability than DL models like LSTM, even when FDIAs are present. Scaling, pulse attacks, and a mixed attack template significantly degrade the accuracy and recall of these models, particularly LSTM, which suffers the most from pulse attacks. To mitigate these effects, two defense strategies were implemented: an autoencoder-enabled LSTM (AE-LSTM) and adversarial training (AT). The AE-LSTM improved stability by reconstructing corrupted inputs and reducing variability under attack, but could not fully restore accuracy in high-intensity or mixed scenarios. In contrast, the adversarially trained models, particularly AT-XGBoost, demonstrated superior robustness, maintaining accuracy above 98.9% even under severe attack conditions. This study further bridges the gap of existing research by experimentally revealing how cyber-attacks impact PdM performance and by introducing a practical adversarial defense strategy to counter these effects. Moreover, the proposed SCADA data preprocessing pipeline enhances the reliability of fault prediction through well-structured data integration, labeling, and feature refinement. These findings highlight the severity of cyber-attacks and the superiority of ML models over DL models, particularly XGBoost, due to its ability to simplify the training process and to minimize the complexity of optimization. Overall, this study provides a unified framework that bridges the gap between data preprocessing, cyber-resilience analysis, and secure PdM implementation for WTs. Future work may explore adversarial attacks, advanced cybersecurity mechanisms, and hybrid modeling approaches to further enhance the robustness of PdM systems in the face of evolving cyber threats.

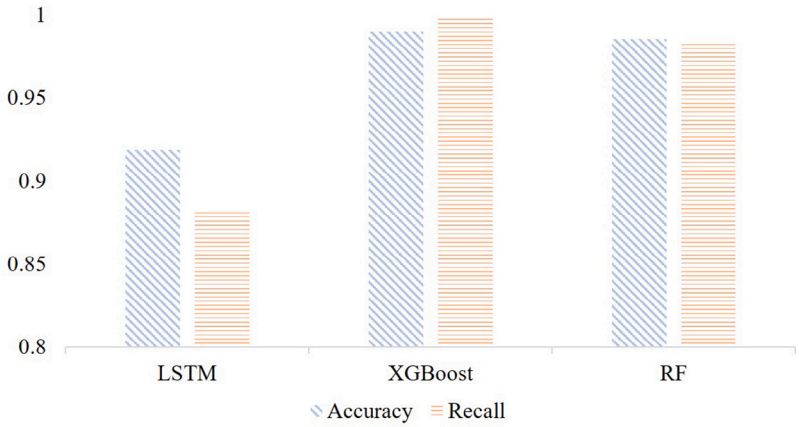

Fig. 3. Performance of adversarially trained models under random attacks.

CRediT authorship contribution statement

Animesh Sarkar Tusher: Writing – original draft, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Md. Abdur Rahman:** Writing – original draft, Resources, Methodology, Conceptualization. **Md. Rashidul Islam:** Writing – review & editing, Supervision, Conceptualization. **Md. Arafat Hossain:** Writing – original draft, Resources. **Adnan Anwar:** Writing – review & editing, Supervision. **M.J. Hossain:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The raw dataset employed in this study was obtained from an open-access data repository, and the preprocessed version of the dataset will be made available upon reasonable request to the authors. Link https://www.kaggle.com/datasets/wasuratme96/iiot-data-of-wind-turbine/.

References

[1] Raja HA, Kudelina K, Asad B, Vaimann T, Rassõlkin A, Kallaste A. Development and utilization of synthetic signals for fault diagnostics of electrical machines. IEEE J Emerg Sel Top Ind Electron 2024;5(4):1447–54. http://dx.doi.org/10.1109/JESTIE.2024.3395650.

[2] Zhansheng L, Jiarong Z, Qingwen Z, Ruilong X. Advances and trends in intelligent maintenance for wind turbine systems. Sustain Energy Technol Assess 2025;80:104398. http://dx.doi.org/10.1016/j.seta.2025.104398.

[3] Falekas G, Palaiologou I, Karlis A, Antonino-Daviu JA. Condition evaluation of steam turbine generator using minute-interval integrated vibration signals. IEEE J Emerg Sel Top Ind Electron 2023;4(3):836–43. http://dx.doi.org/10.1109/JESTIE.2022.3223312.

[4] Mitikiri SB, Babu KVSM, Dwivedi D, Srinivas VL, Chakraborty P, Yemula PK, et al. Cyber–physical security in EV charging infrastructure: Components, vulnerabilities, and defense strategies. Sustain Energy Technol Assess 2025;81:104435. http://dx.doi.org/10.1016/j.seta.2025.104435.

[5] Badihi H, Jadidi S, Yu Z, Zhang Y, Lu N. Smart cyber-attack diagnosis and mitigation in a wind farm network operator. IEEE Trans Ind Inform 2023;19(9):9468–78. http://dx.doi.org/10.1109/TII.2022.3228686.

[6] Pandit R, Infield D, Dodwell T. Operational variables for improving industrial wind turbine yaw misalignment early fault detection capabilities using data-driven techniques. IEEE Trans Instrum Meas 2021;70:1–8. http://dx.doi.org/10.1109/TIM.2021.3073698.

[7] Islam MR, Hasan J, Islam MR, Kouzani AZ, Mahmud MAP. Transient performance augmentation of DFIG based wind farms by nonlinear control of flux-coupling-type superconducting fault current limiter. IEEE Trans Appl Supercond 2021;31(8):1–5. http://dx.doi.org/10.1109/TASC.2021.3091061.

[8] Xu J, Jiang X, Liao S, Ke D, Sun Y, Yao L, et al. Probabilistic prognosis of wind turbine faults with feature selection and confidence calibration. IEEE Trans Sustain Energy 2024;15(1):52–67. http://dx.doi.org/10.1109/TSTE.2023.3272317.

[9] Khan PW, Yeun CY, Byun YC. Fault detection of wind turbines using SCADA data and genetic algorithm-based ensemble learning. Eng Fail Anal 2023;148:107209. http://dx.doi.org/10.1016/j.engfailanal.2023.107209.

[10] Su X, Deng C, Shan Y, Shahnia F, Fu Y, Dong Z. Fault diagnosis based on interpretable convolutional temporal-spatial attention network for offshore wind turbines. J Mod Power Syst Clean Energy 2024;12(5):1459–71. http://dx.doi.org/10.35833/MPCE.2023.000606.

[11] Fu Y, Wang S, Jia F, Zhou Q, Ge X. Two-stage cascaded high-precision early warning of wind turbine faults based on machine learning and data graphization. J Electr Eng Technol 2024;19(3):1919–31. http://dx.doi.org/10.1007/s42835-023-01677-8.

[12] Yakupoglu H, Gözde H, Cengiz Taplamacioglu M. Online noise-adaptive Kalman filter integrated novel autoencoder for multi-fault detection and early warning of wind turbines. Measurement 2025;256:118538. http://dx.doi.org/10.1016/j.measurement.2025.118538.

[13] Lin K-C, Hsu J-Y, Wang H-W, Chen M-Y. Early fault prediction for wind turbines based on deep learning. Sustain Energy Technol Assess 2024;64:103684. http://dx.doi.org/10.1016/j.seta.2024.103684.

[14] Liu J, Wang X, Wu S, Wan L, Xie F. Wind turbine fault detection based on deep residual networks. Expert Syst Appl 2023;213:119102. http://dx.doi.org/10.1016/j.eswa.2022.119102.

[15] Ogaili AAF, Hamzah MN, Jaber AA. Enhanced fault detection of wind turbine using extreme gradient boosting technique based on nonstationary vibration analysis. J Fail Anal Prev 2024;24(2):877–95. http://dx.doi.org/10.1007/s11668-024-01894-x.

[16] Allal Z, Noura HN, Vernier F, Salman O, Chahine K. Wind turbine fault detection and identification using a two-tier machine learning framework. Intell Syst Appl 2024;22:200372. http://dx.doi.org/10.1016/j.iswa.2024.200372.

[17] Lu Q, Ye W, Yin L. Parallel multiple CNNs with temporal predictions for wind turbine blade cracking early fault detection. IEEE Trans Instrum Meas 2024;73:1–11. http://dx.doi.org/10.1109/TIM.2024.3370786.

[18] Maldonado-Correa J, Torres-Cabrera J, Martín-Martínez S, Artigao E, Gómez-Lázaro E. Wind turbine fault detection based on the transformer model using SCADA data. Eng Fail Anal 2024;162:108354. http://dx.doi.org/10.1016/j.engfailanal.2024.108354.

[19] Zhu Y, Xie B, Wang A, Qian Z. Wind turbine fault detection and identification via self-attention-based dynamic graph representation learning and variable-level normalizing flow. Reliab Eng Syst Saf 2025;253:110554. http://dx.doi.org/10.1016/j.ress.2024.110554.

[20] Hu W, Jiao Q, Liu H, Wang K, Jiang Z, Wu J, et al. A transferable diagnosis method with incipient fault detection for a digital twin of wind turbine. Digit Eng 2024;1:100001. http://dx.doi.org/10.1016/j.dte.2024.100001.

[21] Song F, Han Y, William Heath A, Hou M. Structural damage detection of floating offshore wind turbine blades based on Conv1d-GRU-MHA network. Eng Fail Anal 2024;166:108896. http://dx.doi.org/10.1016/j.engfailanal.2024.108896.

[22] Wang Z, Jiang X, Xu Z, Cai C, Wang X, Xu J, et al. Early anomaly detection of wind turbine gearbox based on SLFormer neural network. Ocean Eng 2024;311:118925. http://dx.doi.org/10.1016/j.oceaneng.2024.118925.

[23] Qin S, Tao J, Zhao Z. Fault diagnosis of wind turbine pitch system based on LSTM with multi-channel attention mechanism. Energy Rep 2023;10:4087–96. http://dx.doi.org/10.1016/j.egyr.2023.10.076.

[24] Wang X, Zhu H, Luo X, Guan X. Data-driven-based detection and localization framework against false data injection attacks in DC microgrids. IEEE Internet Things J 2025;12(17):36079–93. http://dx.doi.org/10.1109/JIOT.2025.3579915.

[25] Tatipatri N, Arun SL. A comprehensive review on cyber-attacks in power systems: Impact analysis, detection, and cyber security. IEEE Access 2024;12:18147–67. http://dx.doi.org/10.1109/ACCESS.2024.3361039.

[26] Manandhar K, Cao X, Hu F, Liu Y. Detection of faults and attacks including false data injection attack in smart grid using Kalman filter. IEEE Trans Control Netw Syst 2014;1(4):370–9. http://dx.doi.org/10.1109/TCNS.2014.2357531.

[27] Liu Y, Wu Z, Wang X. Research on fault diagnosis of wind turbine based on SCADA data. IEEE Access 2020;8:185557–69. http://dx.doi.org/10.1109/ACCESS.2020.3029435.

[28] Rahman MA, Islam MR, Hossain MA, Rana MS, Hossain MJ, Gray EM. Resiliency of forecasting methods in different application areas of smart grids: A review and future prospects. Eng Appl Artif Intell 2024;135:108785. http://dx.doi.org/10.1016/j.engappai.2024.108785.

[29] Tusher AS, Rahman MA, Islam MR, Hossain MJ. Adversarial training-based robust lifetime prediction system for power transformers. Electr Power Syst Res 2024;231:110351. http://dx.doi.org/10.1016/j.epsr.2024.110351.

[30] Cui M, Wang J, Yue M. Machine learning-based anomaly detection for load forecasting under cyberattacks. IEEE Trans Smart Grid 2019;10(5):5724–34. http://dx.doi.org/10.1109/TSG.2018.2890809.

[31] Lin K-C, Hsu J-Y, Wang H-W, Chen M-Y. Early fault prediction for wind turbines based on deep learning. Sustain Energy Technol Assess 2024;64:103684. http://dx.doi.org/10.1016/j.seta.2024.103684.

[32] Tusher AS, Rahman MA, Islam MR, Hossain MJ. A comparative analysis of different transmission line fault detectors and classifiers during normal conditions and cyber-attacks. J Eng 2024;2024(7):e12412. http://dx.doi.org/10.1049/tje2.12412.

[33] Kaggle. IIOT data of wind turbine. 2024, URL https://www.kaggle.com/datasets/wasuratme96/iiot-data-of-wind-turbine. [Accessed: 29 October 2024].

[34] Safavian S, Landgrebe D. A survey of decision tree classifier methodology. IEEE Trans Syst Man Cybern 1991;21(3):660–74. http://dx.doi.org/10.1109/21.97458.

[35] Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Statist 2001;29(5):1189–232. http://dx.doi.org/10.1214/aos/1013203451, .

[36] Trizoglou P, Liu X, Lin Z. Fault detection by an ensemble framework of Extreme Gradient Boosting (XGBoost) in the operation of offshore wind turbines. Renew Energy 2021;179:945–62. http://dx.doi.org/10.1016/j.renene.2021.07.085.